Opening the black box: how we explain and check large language models

Chatbots like GPT, Claude, Gemini, and Llama make decisions no one can fully read. This map shows why that is hard, what the four biggest labs are doing about it, and how testing turns an explanation into something you can trust.

Bindu Bhargava Reddy Chintam
AIML-501 Model Development
Indiana Wesleyan University
Sources checked November 2026

What explainable AI means, and why it matters

Explainable AI is the work of showing people why a model gave an answer, in a way they can understand and check.

Trust
Doctors, judges, and customers need reasons before they act on an answer.

Safety
Engineers can only fix a failure they can see.

Fairness
Explanations reveal when a model leans on race, gender, or dialect.

The law
The EU AI Act gives people affected by high-risk AI decisions a right to a clear explanation.

Four promises a good explanation should keep

Explanation	Meaningful	Accurate	Knows its limits
It gives reasons or evidence for the output.	Its intended reader can understand it.	It matches how the model really decided.	It flags when the model is out of its depth.

Based on NIST's four principles of explainable AI (Phillips et al., 2021). Most of today's problems break the third promise.

Five reasons large language models are hard to explain

The first two come from inside the model, the third from its training data, and the last two from the people and rules around it.

1. Inside the model
Too big and tangled
Knowledge is spread across billions of numbers, and one neuron can stand for several unrelated ideas at once.

34 million
concepts Anthropic found in one mid-size Claude model, and still not all of them

2. Inside the model
Explanations that are not the real reason
When a model explains itself, the words can be a story told after the fact.

up to 36%
accuracy drop from a hidden bias that the model never mentioned in its reasoning

3. Training data
Hidden data, inherited bias
Models learn from data the public cannot inspect, including its stereotypes.

Covert bias
GPT-4 judged dialect speakers more harshly, even though its direct answers about race sounded positive

4. People
One answer, many audiences
A patient, a doctor, an engineer, and a regulator each need a different kind of explanation.

Meaningful
is one of NIST's four principles: the reader must understand it

5. Rules and business
Secrecy and new laws
Closed models protect trade secrets, while laws now demand explanations for high-risk decisions.

Article 86
of the EU AI Act: a right to a clear explanation of high-risk AI decisions

Which tools answer which problem

Each column is a family of techniques, with real examples from OpenAI, Anthropic, Google, and Meta. Each row is a challenge from above. A filled dot means the technique was built to address it.

Challenge Filled dot: main tool. Half dot: partial help.	Look inside the model *What is happening in the network?*	Ask the model to show its work *What does the model say it did?*	Trace answers back to inputs *Which words or features drove this answer?*	Document and open up *What can outsiders check?*
1. Too big and tangled Inside the model	● Finds the concepts and circuits directly	◐ Readable steps give a window, not the full picture	◐ Shows which inputs mattered, not why	Not its job
2. Explanations that are not the real reason Inside the model	● Checks the stated reason against what really fired inside	● Monitors and tests the stated reasoning	◐ Compares the claimed reason with input importance	Not its job
3. Hidden data, inherited bias Training data	◐ Can surface concepts tied to bias	Not its job	● Reveals reliance on names, dialect, or gender words	● Data and bias tests reported in cards
4. One answer, many audiences People	Not its job	● Plain-language reasons for everyday users	◐ Highlighted words help experts review a case	● Cards written for decision makers
5. Secrecy and new laws Rules and business	◐ Gives evidence regulators may later expect	◐ Reasoning logs support audits	◐ Per-decision reasons for affected people	● Disclosure and outside access meet transparency duties

Look inside the model
Anthropic Finds millions of concepts in Claude and traces how they connect into step-by-step "circuits"
OpenAI Maps 16 million concepts in GPT-4; uses GPT-4 to describe GPT-2's neurons
Google Gemma Scope: free concept maps for every layer of its smaller open Gemma 2 models

Ask the model to show its work
OpenAI Lets a second model read reasoning steps to catch cheating
Anthropic Tests whether Claude's reasoning admits the hints it used
All four Joint 2025 paper urges labs to keep reasoning readable

Trace answers back to inputs
Meta Captum toolkit scores how much each input mattered
Research SHAP values and attention maps; attention alone can mislead

Document and open up
Google Model cards: a standard fact sheet for each model
OpenAI, Anthropic System cards with safety test results
Meta, Google Open model weights that any researcher can study

How validation and metrics turn an explanation into trust

An explanation is only a claim until it is tested. These four checks, run in order and repeated, are how labs and users decide whether a model, and its explanation, can be relied on.

Check 1
Score it on new data
Measure precision, recall, and calibration on data the model never saw, broken out by group.

3 of 42
cancers missed by a 96%-accurate model in my Workshop 4 test, which accuracy alone hid

Check 2
Attack it
Red-teamers and test suites look for unsafe answers, bias, and wrong refusals.

~6%
of 250 safe prompts that GPT-4 still refused in one test suite

Check 3
Test the explanation itself
Plant a hint or change an internal concept and see if the explanation and the answer change together.

25% / 39%
how often Claude 3.7 Sonnet and DeepSeek R1 admitted using a planted hint

Check 4
Watch it after launch
Monitor outputs and reasoning in real use, and publish results in model and system cards.

1 in 4
prompts where Anthropic's best tracing tool gave a clear answer, so monitoring still matters

A failed check sends the team back to the data, the model, or the explanation, so the loop never really ends.

What this means for anyone using AI

A model's own explanation is a clue, not proof.
When tested, reasoning models admitted using a planted hint only 25% to 39% of the time.

Looking inside is real but early.
Labs can now find millions of concepts inside models, yet the best tools give clear answers for only about a quarter of the prompts tried.

Trust comes from testing, not from words.
Ask for the metrics: recall and precision on your own data, bias tests, red-team results, and how the explanation itself was checked.

Sources. Baker et al. (2025); Bills et al. (2023); Chen et al. (2025); Gao et al. (2024); Hofmann et al. (2024); Jain & Wallace (2019); Kokhlikyan et al. (2020); Korbak et al. (2025); Lieberum et al. (2024); Lindsey et al. (2025); Lundberg & Lee (2017); Mitchell et al. (2019); Phillips et al. (2021); Regulation (EU) 2024/1689, Art. 86; Röttger et al. (2024); Templeton et al. (2024); Turpin et al. (2023). Full APA 7 references are in the accompanying document.